SoundGuides: Adapting Continuous Auditory Feedback to Users

Jules Françoise
SMTS Lab, Ircam-CNRS-UPMC
1, Place Igor Stravinsky 75004
Paris, France.
francoise@ircam.fr

Olivier Chapuis
LRI – Univ Paris-Sud, CNRS
Inria & Université Paris-Saclay
F-91405 Orsay, France
chapuis@lri.fr

Sylvain Hanneton
LPP - UMR 8242
Université Paris-Descartes, 45
rue des Saints-Pères, 75006
Paris, France
sylvain.hanneton@parisdescartes.fr

Frédéric Bevilacqua
SMTS Lab, Ircam-CNRS-UPMC
1, Place Igor Stravinsky 75004
Paris, France.
bevilacqua@ircam.fr

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). Copyright is held by the author/owner(s).
CHI'16 Extended Abstracts, May 7–12, 2016, San Jose, CA, USA.
ACM 978-1-4503-4082-3/16/05.
http://dx.doi.org/10.1145/2851581.2892420

Abstract
We introduce *SoundGuides*, a user adaptable tool for auditory feedback on movement. The system is based on a interactive machine learning approach, where both gestures and sounds are first conjointly designed and conjointly learned by the system. The system can then automatically adapt the auditory feedback to any new user, taking into account the particular way each user performs a given gesture. *SoundGuides* is suitable for the design of continuous auditory feedback aimed at guiding users' movements and helping them to perform a specific movement consistently over time. Applications span from movement-based interaction techniques to auditory-guided rehabilitation. We first describe our system and report a study that demonstrates a 'stabilizing effect' of our adaptive auditory feedback method.

Author Keywords
Movement; Sound; 3D Gestures; Auditory feedback; User adaptation

ACM Classification Keywords
H.5.2. [Information Interfaces and Presentation (e.g. HCI)]: User Interfaces; H.5.5. [Information Interfaces and Presentation (e.g. HCI)]: Sound and Music Computing

Introduction

Movement is central in so called embodied interaction techniques, which interest is growing in Human-Computer Interaction (HCI). As previously reported in the HCI community, the use of movement-based interfaces often requires learning, memorization and expertise [1, 2, 4]. Therefore, when designing interaction with movement-based interfaces, one must carefully consider how users adapt or learn to perform new gestures. Designing continuous feedback informing users in real-time on the quality of their performance has the potential to support such a learning process. We focus here on continuous auditory feedback, which is particularly suitable for interaction paradigms where the feedback cannot be given visually. Moreover, several works point our the potential of auditory continuous feedback for either learning or improving gesture performance [6, 9, 18, 20].

Nonetheless, *continuous* and in particular *gesture-synchronized* feedback have been only partially studied in the HCI community. First, programming such continuous feedback remains generally challenging, especially when relationships between specific — and potentially complex — gestures and customized auditory feedback are sought. Second, beyond simplistic movement-sound relationships, users often have difficulties to comprehend the auditory feedback, which generally result in difficult user adaptation.

In this paper, we describe a system allowing designers to easily synchronize movement parameters to sound synthesis, without any programming. Most importantly, the proposed approach, we call *SoundGuides*, allows for directly adapting the auditory feedback to the idiosyncrasies of a new user performing the gesture. We describe in this note the *SoundGuides* system, available as open-software, along with its implementation with a common movement tracking system. We also report a study that highlights a 'stabilizing effect' of our adaptive auditory feedback method.

Related Work

Gesture Learning with Visual Feedback

As tangible and gestural interfaces become ubiquitous, users must constantly adapt to new interaction techniques that require learning and mastering new gestures. Although studies support that user-defined gestures are easier to recall and execute [15, 24], they can be challenging due to misconceptions of the recognizer's abilities [17]. For robust recognition, predefined gesture sets are the most widespread, which led to the development of a thread of HCI concerned with providing users with novel means to learn such gesture sets.

Several methods use onscreen visual display to guide gesture interaction. Octopocus is a dynamic guide combining feedforward and feedback mechanism for helping users to "learn, execute and remember gesture sets" [4]. Dynamic guides significantly reduce the input time with surface gestures or strokes compared to help menus and hierarchical marking menus [2, 4].

Most approaches focus on multi-touch gestures on two-dimensional surfaces devices where visual feedback can be co-located and situated. We consider in this note the case of three-dimensional mid-air gestures that have become essential in full-body interaction, for example with large displays [16]. In this case, situated visual feedback is difficult to implement and might add a heavy cognitive load. As an alternative, we investigate the use of sound as a feedback modality, and propose to study how continuous auditory feedback can support gesture performance.

Motor Learning with Auditory Feedback

While vision has long been the primary feedback modality in motor control studies, a recent thread of sensori-motor learning research investigates audition as an extrinsic feed-

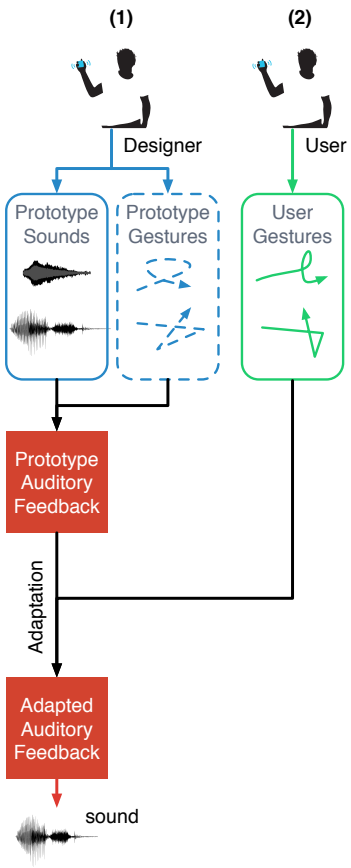

Figure 1: Principle of *SoundGuides*: (1) Designers elaborate a set of prototype gestures with customized synchronous sound; (2) end users record their own gestures so that the system adapts the auditory feedback.

back modality. Auditory perception has a lower cognitive load and faster response than visual feedback [3], yet it can carry rich information content. Using interactive auditory feedback for movement learning has many applications such as HCI [18], motor rehabilitation [20], or sport where it aims to improve performance and accuracy [9].

According to Anderson et al. [1], learnability involves two factors: the cognitive mapping between gestures and actions (associative learning), and the ability to perform a gesture. Because it is already well-known that auditory feedback can help associative learning [14], in this paper we focus on the potential of auditory feedback for improving the performance of arbitrary gestures.

Often, direct mapping strategies are used for sonifying physical quantities. Reviewing 179 publications related to motion sonification, Dubus et al. [8] highlight that direct mapping strategies and simple sound synthesis prevail — many works use pure tones varying in pitch, amplitude and panning, often directly driven by position or velocity. However, as noted by Sigrist et al. [23], a major drawback of direct sonification is the difficulty of specifying the 'target' movements in high expertise applications such as sport. Moreover, the use of basic sound synthesis can be ineffective in mid- to long-term learning tasks: practicing with unpleasant sound feedback can even degrade performance. We propose to use rich sound textures with a focus on timbral variations. Our system is flexible enough to allow designers and users to easily customize the sounds.

While most approaches to movement auditory feedback use hard-coded mapping strategies, recent research in HCI and New Interfaces for Musical Expression (NIME) address end-user design of mapping strategies through interactive machine learning. Systems such as the Wekinator [10] allow for rapid prototyping of classifiers and regression models. The *SoundGuides* system is based on advanced techniques that allow for on continuous recognition and for synchronizing the temporal evolution of gestures in real-time with sound synthesis [5, 7, 13]. *SoundGuides* also enables fast adaptation to the user's movement from a single example.

SoundGuides

SoundGuides is a generic system for providing users with a continuous sound feedback that adapts to their particular way of performing a gesture. While most guides for gesture learning rely upon a fixed feedback strategy, we propose a flexible system that allows both designers and end users to craft their own gestures.

Principle

SoundGuides implements a 2-steps process, implying first the designer and then the user, as described in Figure 1. From the designer's perspective, the first step consists in the elaboration of gestures, each of them being synchronized to sound examples. We particularly encourage the design of sounds which parameters (e.g. intensity, pitch, timbre, etc.) evolves consistently with the gesture's dynamics. Designers can customize the feedback easily by uploading their own recordings to the system.

The sound synthesis, based on corpus-based concatenative synthesis [21], has the ability to play and transform any audio recording in real-time through intuitive control parameters. In particular, it is possible to selectively choose which sound characteristics should be modified or kept constant. The design of the sound examples can be realized using CataRT's graphical interface that make sound design easy through the edition of 2D trajectories, as fully described in [22]. Once the sound and gesture examples have been chosen, we use an Interactive Machine Learn-

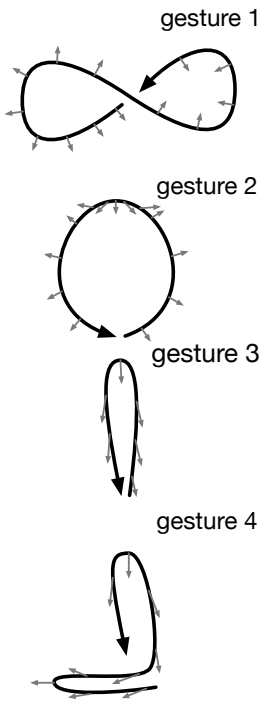

Figure 2: Graphical representation of the four gestures used in the experiment. Black arrow indicate hand position trajectory, thin grey arrows indicate palm direction.

ing approach to learn the mapping between motion parameters and sound synthesis. For this, the designer typically records few examples of gestures while listening to the sounds examples [12]. Importantly, this step can be repeated and allows for iterating over the design process.

Step 2: User Adaptation
While the designer can provide a set of 'prototype' gestures with their associated mapping and auditory feedback strategy, *SoundGuides* allows end users to directly record their own version of the prototype gestures. The system can then adapt to the idiosyncrasies of their performance. In particular, the machine learning model of the motion-sound mapping is trained with the user demonstration of the prototype gesture. This approach allow designers to specify a gesture set, suited for a particular recognition system of interaction technique, but it also gives users the possibility to adapt the feedback to the particularities of their performance.

SoundGuides is thus a customizable tool for designing adaptive continuous auditory feedback on movement. Both gestures and sounds can be edited by designers without requiring programming.[1]

Implementation
SoundGuides is implemented as a modular architecture within Cycling'74 Max and is available as free software. In particular, the mapping between motion parameters and sound descriptors is learned using Gaussian Mixture Regression (GMR), a regression method based on Gaussian Mixture Models. GMR learns a smooth and possibly non-linear mapping function from the set of sound and gesture examples provided by the user.[2] Our system can be easily adapted to various input devices such as positional tracking or inertial sensors. In this paper, we present a concrete implementation that uses 3D hand tracking.

Evaluation

We evaluated *SoundGuides* as a user-adaptive tool for providing continuous auditory feedback. Our hypothesis is that the adaptive auditory feedback can stabilize the gesture, even after a short learning phase. By stabilization, we refer to minimizing gesture variations in both time and space.

We designed 4 gestures and their associated sounds (see Figure 2). Participants were asked to imitate a set of 3D gestures from an audio-visual demonstration, and record several executions of each gesture, either without feedback, or with a *SoundGuides* adapted to their own initial performance. We evaluate the intra-users gesture variations with regards to the presence of the continuous auditory feedback.

Participants
We recruited 12 participants, gender-balanced, aged from 19 to 47 (mean=26.9, SD=9.2). All participants were right-handed, the experiment was exclusively performed with the right hand.

Apparatus
The experiment used the 3D velocity of the hand (cartesian coordinates) from the skeleton captured with a Leap Motion. The interface, developed with Cycling'74 Max and running on an Apple MacBook Pro, integrated movement acquisition, mapping, sound synthesis, and the GUI elements necessary to the experiment.

[1] Supplementary material (including a video demonstration) is available online: http://julesfrancoise.com/soundguides

[2] More detail of the algorithmic part and its evaluation can be found in [11]. Our GMR algorithm is part of the open-source library XMM: https://github.com/Ircam-RnD/xmm.

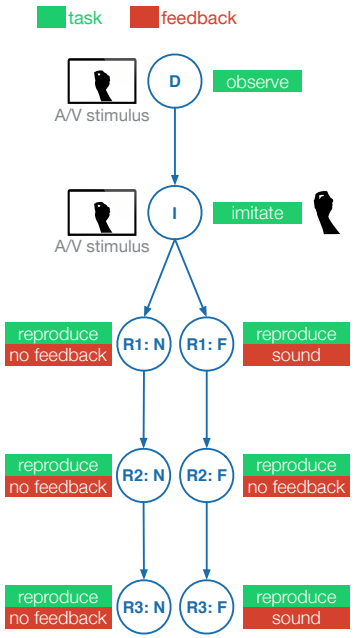

Figure 3: Protocol of one block of the experiment (F=Sound Feedback, N=No Feedback). All participants performed four blocks, one for each gesture, half of them in condition F, the other half in condition N.

Tasks

After on a series of pilot studies, we designed four gestures associated with specific sounds. The sounds were 3 to 4 seconds long, based on wind and water recording. They all presented continuous variations of the audio parameters Loudness and Spectral Centroid [19]. We made an audio-visual recording of one execution of each gesture we call the 'reference gestures', with their associated sound. The participants were asked to reproduce the reference gestures 'as accurately as possible' and to record several repetitions, trying to be 'as consistent as possible'. To facilitate the understanding of the gesture's shape and dynamics, it was videotaped from the viewpoint of the performer — to maximize the correspondence between the participants viewpoint and the reference movement.

Procedure

Feedback was the primary factor with two levels: *Feedback* (**F**) and *No Feedback* (**N**). Each participant performed two gestures with auditory feedback (**F**) (in specific blocks) and the two other gestures without any feedback (silence in all blocks) (**N**). The six possible associations between gestures and conditions were balanced across participants, and the order of presentation of the gestures was randomized, under the constraint that two gestures with the same condition cannot follow each other.

Figure 3 summarizes the procedure for one gesture. The first block (**D**) is the *Demonstration*: the audio-visual recording is played 10 times without interruption. In the second block (**I**), participants must *imitate* the gesture while watching the demonstration, again 10 times. At the end of block **I**, the recorded gestures are used to adapt the mapping to the participant's idiosyncrasies. The experiment continues with 3 recording blocks containing each a series of 10 executions (**R1–R3**), with a pause between each execution.

In condition **N**, all 3 recording blocks are performed without feedback. In condition **F**, the movement is sonified in the first and last recording blocks (**R1^{S}**, **R3^{S}**), and no feedback is provided in the middle recording block (**R2**). A 30 seconds break is imposed between each block to avoid fatigue.

Analysis

To investigate gesture stability, we must calculate distances between gestures. The choice of a distance measure follow several constraints specific to our case: sensitive to both timing variations (time compression/stretching) and amplitude variation (scaling). Nevertheless, it must remain invariant to constant time delays. We rule out the euclidean distance, that does not allow any possible delay or temporal variation between the target and test gestures, and Dynamic Time Warping (DTW) that is, by design, not sensitive enough to temporal variations.

We used a constrained correlation distance, defined as the minimal euclidean distance between two gestures with varying delays:

$$d(\boldsymbol{g}_{ref}, \boldsymbol{g}_{test}) = \min_{\tau \in [-\Delta, \Delta]} \frac{1}{T_\tau} \sum_{t=1}^{T_\tau} \|\boldsymbol{g}_{ref}(t) - \boldsymbol{g}_{test}(t-\tau)\|$$

where $\|\boldsymbol{a} - \boldsymbol{b}\|$ is the euclidean distance between frames $\boldsymbol{a}$ and $\boldsymbol{b}$, T_τ is the number of overlapping frames between the two sequences with delay τ, Δ is the maximum authorized delay.

Results

First, we evaluate the inter-user variations for each gesture. The *standard deviation over mean* ratio is 0.33, 0.25, 0.24, 0.29 for the four gestures, respectively. This clearly demonstrates a large variability among users, even for relatively simple gestures that were asked to be reproduced as accurately as possible. We also note that gesture 1 and 4 are the ones that exhibit the most variability. This demonstrates the importance to take into account user variability when

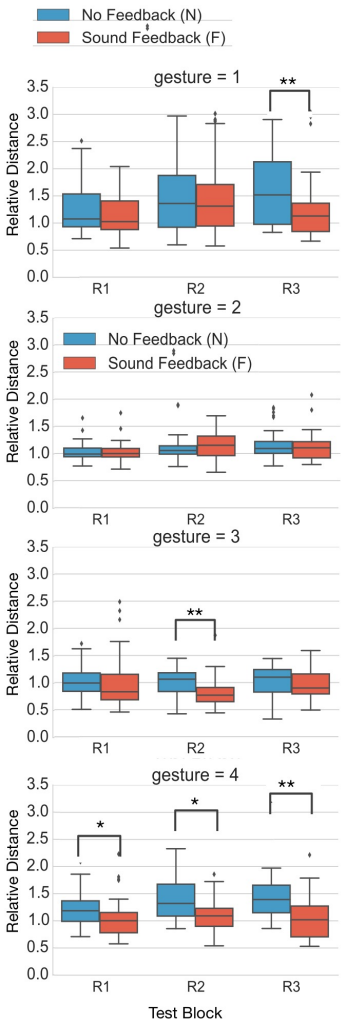

Figure 4: Normalized distance to the 'reference gesture' for each gesture and each block of the experiment, across participants and trials. Stars indicate statistical significance under p<0.01 and p<0.001.

limited learning time is allocated to the users. From now on, we report only on intra-user variability by normalizing the distances of user gestures in recording blocks by their corresponding average distance in the imitation block.

Figure 4 details the distances between trials and reference gestures across participants and trials, for each of the four gestures. We examine whether the normalized distance depends on the **N** and **F** conditions. We tested for statistical significance using non-parametric Mann-Whitney U tests for each gesture and each recording block across participants. Gestures 1 and 4 exhibit clear differences between feedback conditions. In condition **N**, the distance to the reference gesture tends to increase along recording blocks. On the contrary, the distance remains approximately constant and close to 1 across recording blocks when the gestures are performed with the help of auditory feedback. This highlights a stabilizing effect of the auditory feedback.

Specifically, for gesture 1, we found a significant difference in the distances in recording block **R3** where the medians for conditions **N** and **F** are respectively 1.52 and 1.13 (The mean ranks of Group **F** and Group **N** were 45 and 64, respectively; $U = 912, Z = -3.18, p < 0.001, r = 0.31$). For gesture 4, the distance was found lower in condition **F** than in condition **N** for all three recording blocks, under $p < 0.01$. This difference is all the more important in **R3**, where the median distance in conditions **N** and **F** is respectively 1.39 and 1.02 (The mean ranks of Group **F** and Group **N** were 38 and 69, respectively; $U = 643, Z = -5.06, p < 0.001, r = 0.48$).

For gestures 2 and 3, we found much lower differences differences over time in both conditions. In this case, the gesture seems sufficiently stable even without any feedback, and the influence of auditory feedback is not significant. Nevertheless, for gesture 3, we found a significant difference between the condition **F** and **N** (The mean ranks of Group **F** and Group **N** were 43 and 70, respectively; $U = 901, Z = -4.14, p < 0.001, r = 0.38$). This might be explained by an after-effect between **R1** (with auditory feedback) and **R2** (where the auditory feedback is removed).

Discussion and Conclusion

We described *SoundGuides*, a user-adaptive tool for providing continuous auditory feedback. The system was evaluated in a task where users can record their own version of prototypical gestures. The results indicate that the continuous auditory feedback can help users minimize the gestures variations. This was found to be significant when the gestures are not performed consistently over time without any feedback. We also found that the efficiency of the auditory feedback highly depends on the gesture itself, which could be related to the gesture difficulty. More studies are needed to answer this important but vast question. In particular, We plan to further correlate the impact of the performance difficulty with the effect of the auditory feedback.

Overall, we believe that our results show the promise of *SoundGuides*, opening novel opportunities to build movement-based interfaces. For example, gesture recognition could be improved by providing users with such auditory feedback, since the reproducibility of movement is essential in such applications. *SoundGuides* can also find applications in rehabilitation where movement can be guided by auditory feedback. In such applications, the adaptation to the user idiosyncrasies is particularly important since each patient might suffer from severe movement limitations.

Acknowledgments

This work is funded by the LEGOS project (ANR Grant 31639884) and by the Rapid-Mix EU project (H2020-ICT-2014-1 Project ID 644862).

References

[1] Fraser Anderson and Walter F Bischof. 2013. Learning and performance with gesture guides. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, Paris, France, 1109–1118. DOI: http://dx.doi.org/10.1145/2470654.2466143

[2] Caroline Appert and Shumin Zhai. 2009. Using Strokes As Command Shortcuts: Cognitive Benefits and Toolkit Support. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, Boston, USA, 2289–2298. DOI: http://dx.doi.org/10.1145/1518701.1519052

[3] Yoram Baram and Ariel Miller. 2007. Auditory feedback control for improvement of gait in patients with Multiple Sclerosis. *Journal of the Neurological Sciences* 254 (2007), 90–94. DOI: http://dx.doi.org/10.1016/j.jns.2007.01.003

[4] Olivier Bau and Wendy E Mackay. 2008. OctoPocus: A Dynamic Guide for Learning Gesture-based Command Sets. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology (UIST '08)*. ACM, 37–46. DOI: http://dx.doi.org/10.1145/1449715.1449724

[5] Frédéric Bevilacqua, Bruno Zamborlin, Anthony Sypniewski, Norbert Schnell, Fabrice Guédy, and Nicolas Rasamimanana. 2010. Continuous realtime gesture following and recognition. *Gesture in Embodied Communication and Human-Computer Interaction* (2010), 73–84. DOI: http://dx.doi.org/10.1007/978-3-642-12553-9_7

[6] Eric O. Boyer, Quentin Pyanet, Sylvain Hanneton, and Frédéric Bevilacqua. 2014. Learning Movement Kinematics with a Targeted Sound. In *Lecture Notes in Computer Science*. Vol. 8905. Springer Verlag, 218–233.

[7] Baptiste Caramiaux, Nicola Montecchio, Atau Tanaka, and Frédéric Bevilacqua. 2014. Adaptive Gesture Recognition with Variation Estimation for Interactive Systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 4, 4 (2014), 18:1—-18:34. DOI: http://dx.doi.org/10.1145/2643204

[8] Gaël Dubus and Roberto Bresin. 2013. A systematic review of mapping strategies for the sonification of physical quantities. *PloS one* 8, 12 (Jan. 2013), e82491. DOI: http://dx.doi.org/10.1371/journal.pone.0082491

[9] Alfred Effenberg, Ursula Fehse, and Andreas Weber. 2011. Movement Sonification: Audiovisual benefits on motor learning. *BIO Web of Conferences* 1 (Dec. 2011). DOI: http://dx.doi.org/10.1051/bioconf/20110100022

[10] Rebecca Fiebrink, Perry R. Cook, and Dan Trueman. 2011. Human model evaluation in interactive supervised learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*. ACM, Vancouver, BC, Canada, 147. DOI: http://dx.doi.org/10.1145/1978942.1978965

[11] Jules Françoise. 2015. *Motion-Sound Mapping by Demonstration*. PhD Dissertation. Université Pierre et Marie Curie. http://julesfrancoise.com/phdthesis

[12] Jules Françoise, Norbert Schnell, and Frédéric Bevilacqua. 2013. A Multimodal Probabilistic Model for Gesture-based Control of Sound Synthesis. In *Proceedings of the 21st ACM international conference on Multimedia (MM'13)*. Barcelona, Spain, 705–708. DOI: http://dx.doi.org/10.1145/2502081.2502184

[13] Jules Françoise, Norbert Schnell, Riccardo Borghesi, and Frédéric Bevilacqua. 2014. Probabilistic Models for Designing Motion and Sound Relationships. In *Proceedings of the 2014 International Conference on New Interfaces for Musical Expression (NIME'14)*. London, UK, 287–292.

[14] William Gaver. 1986. Auditory Icons: Using Sound in Computer Interfaces. *Human-Computer Interaction* 2, 2 (June 1986), 167–177. DOI: http://dx.doi.org/10.1207/s15327051hci0202_3

[15] Miguel A Nacenta, Yemliha Kamber, Yizhou Qiang, and Per Ola Kristensson. 2013. Memorability of Pre-designed and User-defined Gesture Sets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, 1099–1108. DOI: http://dx.doi.org/10.1145/2470654.2466142

[16] Mathieu Nancel, Julie Wagner, Emmanuel Pietriga, Olivier Chapuis, and Wendy Mackay. 2011. Mid-air pan-and-zoom on wall-sized displays. In *Proceedings of the 2011 annual conference on Human factors in computing systems (CHI '11)*. ACM, ACM Press, Vancouver, BC, Canada, 177. DOI: http://dx.doi.org/10.1145/1978942.1978969

[17] Uran Oh and Leah Findlater. 2013. The challenges and potential of end-user gesture customization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. ACM Press, New York, New York, USA, 1129. DOI: http://dx.doi.org/10.1145/2470654.2466145

[18] Uran Oh, Shaun K. Kane, and Leah Findlater. 2013. Follow that sound: using sonification and corrective verbal feedback to teach touchscreen gestures. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS '13*. ACM Press, New York, New York, USA, 1–8. DOI: http://dx.doi.org/10.1145/2513383.2513455

[19] G Peeters. 2004. *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. Technical Report. http://www.citeulike.org/group/1854/article/1562527

[20] Johanna V G Robertson, Thomas Hoellinger, På vel Lindberg, Djamel Bensmail, Sylvain Hanneton, and Agnès Roby-Brami. 2009. Effect of auditory feedback differs according to side of hemiparesis: a comparative pilot study. *Journal of neuroengineering and rehabilitation* 6 (Jan. 2009), 45. DOI: http://dx.doi.org/10.1186/1743-0003-6-45

[21] Diemo Schwarz. 2007. Corpus-based concatenative synthesis. *Signal Processing Magazine, IEEE* 24, 2 (2007), 92–104.

[22] Diemo Schwarz, Grégory Beller, Bruno Verbrugghe, Sam Britton, and others. 2006. Real-time corpus-based concatenative synthesis with catart. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx), Montreal, Canada*. Citeseer, 279–282.

[23] Roland Sigrist, Georg Rauter, Robert Riener, and Peter Wolf. 2013. Augmented visual, auditory, haptic, and multimodal feedback in motor learning: a review. *Psychonomic bulletin & review* 20, 1 (March 2013), 21–53. DOI: http://dx.doi.org/10.3758/s13423-012-0333-8

[24] Jacob O Wobbrock, Meredith Ringel Morris, and Andrew D Wilson. 2009. User-defined Gestures for Surface Computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, 1083–1092. DOI: http://dx.doi.org/10.1145/1518701.1518866